April 2015

ANALYSIS OF NSY-1 IN C. ELEGANS AND PARASITIC NEMATODES USING BIOINFORMATICS

Maria Jeanne Grandoni
Worcester Polytechnic Institute

ANALYSIS OF NSY-1 IN *C. ELEGANS* AND PARASITIC NEMATODES USING BIOINFORMATICS

A Major Qualifying Project Report

Submitted to the Faculty

Of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Bachelor of Science

By

Maria Grandoni

Date:

Approved

Table of Contents

List of Figures

List of Tables

Abstract

Infections caused by parasitic nematodes can become difficult to treat due to nematodes changing their outer surface throughout their life cycle and causing the host immune system to fail to recognize each new surface composition. From the model organism *Caenorhabditis elegans (C. elegans)* it has been determined that the gene *srf-6*, which controls changes in nematode surface composition, may be identical to a *C. elegans* MAPKKK gene known as *nsy-1.* The goal of this project is to identify similar proteins in parasitic nematodes. I used the software package Geneious to compare the *C. elegans nsy-1* protein sequence to corresponding proteins in parasitic nematode databases. Because *nsy-1/srf-6* may affect surface antigen switching in parasitic nematodes, it is a possible target for anti-nematode therapies.

Introduction

Nematode anatomy. The animal phylum Nematoda contains both parasitic roundworms and free living roundworms such as *Caenorhabditis elegans (C. elegans)*. Nematode anatomy is consistent among all nematodes, with a body wall, a concentric tube that surrounds another smaller concentric tube, the intestine. Between these two tubes is a space filled with fluid known as the pseudocoelom (Riddle et al, 1997). This space maintains the shape of the nematode through hydrostatic pressure. The pseudocoelom contains the gonads.

Nematode development. Postembryonic development of all nematodes consists of five stages, including four larval stages (L1-L4) and the adult. The stages are separated from each other by molts in which the nematode completely sheds its old cuticle, revealing the presence of the newly synthesized cuticle of the next stage. The adult stage reproduces without additional molts. All levels of cuticle structure, including protein composition and ultrastructure, differ between the different stages of the cuticle in the model nematode *Caenorhabditis elegans* (Cox Staprans and Edgar 1981).

Nematode lifestyles and life cycles. Nematodes are a very successful phylum, with both free-living and parasitic species present. Free-living parasites spend their entire life cycle living in environments such as soil or water, whereas parasitic nematodes spend one or more stages of the life cycle inside of a host organism. Both vertebrate and invertebrate hosts are parasitized by nematodes, and both plants and animals have parasitic nematode infections. Many parasitic nematodes exhibit a restricted host range which may include only one or several related host species.

The small soil nematode *C. elegans* is a well-studied example of a free-living nematode. This species is the most studied of all nematodes. *C. elegans* has a 3 day generation time, produces 300 offspring per hermaphrodite parent, and has a typical nematode life cycle. Once hatched from eggs, the nematode larvae are at the L1 stage. The L1 molts into an L2, and then makes a decision to continue development by molting into an L3, or molt into a dauer larva, depending on the environmental conditions. The dauer/L3 decision is controlled by environmental signals. If food is plentiful and nematodes are not crowded, the L2 stage larva will become an L3, and if otherwise, the L2 will molt into a dauer larva, which is specifically adapted for surviving for long periods without food. If food becomes plentiful again, the dauer larvae molt into the L4 stage and continue development to adults (Edgar et al, 1982).

The parasitic nematode species *Trichinella spiralis* serves as a useful example of a parasitic nematode lifestyle. *Trichinella* L1 infective larvae can be passed from one host to another by encysting in undercooked meat (CDC, 2012), Digestion of the meat in the host stomach by low pH and pepsin releases the L1 infective larvae, which then are passed to the

intestine. In the intestine, the larvae complete the developmental process and become adults. The adults mate and lay eggs, which hatch in the intestine. The newborn L1 larvae then burrow through the intestinal walls and migrate through the host tissues until they reach muscle. They then enter muscle cells where they encyst, converting the muscle cells into nurse cells which support the dormant stage. When another animal or human eats the uncooked muscle of an infected animal, the life cycle is completed. (Trichinella.org, 2004). Interestingly, the infective stage of *T. spiralis* and many other parasitic species is analogous to the *C. elegans* dauer larva, in that a change in environmental conditions triggers the re-entry of the arrested stage into the developmental cycle.

Nematode phylogeny. Originally nematode phylogeny was primarily based on morphology. With improvements in phylogenetics and classification, nematode phylogeny is now based on small subunit (SSU) rRNA sequences (Blaxter et al 1998 ,De Lay, 2006). The use of SSU rRNA showed the traditional phylogeny based on morphological analysis was incorrect. Originally the nematode phylum was made up of two classes, the Adenophorea and Secernentea (Blaxter et al, 1998). Through the use of SSU rRNA sequence analysis, class Adenophorea could be split into two separate classes, so that today three classes are distinguished, known as Chromadoria, Dorylaimia, and Enoplia (Blaxter 2011, Figure 1).

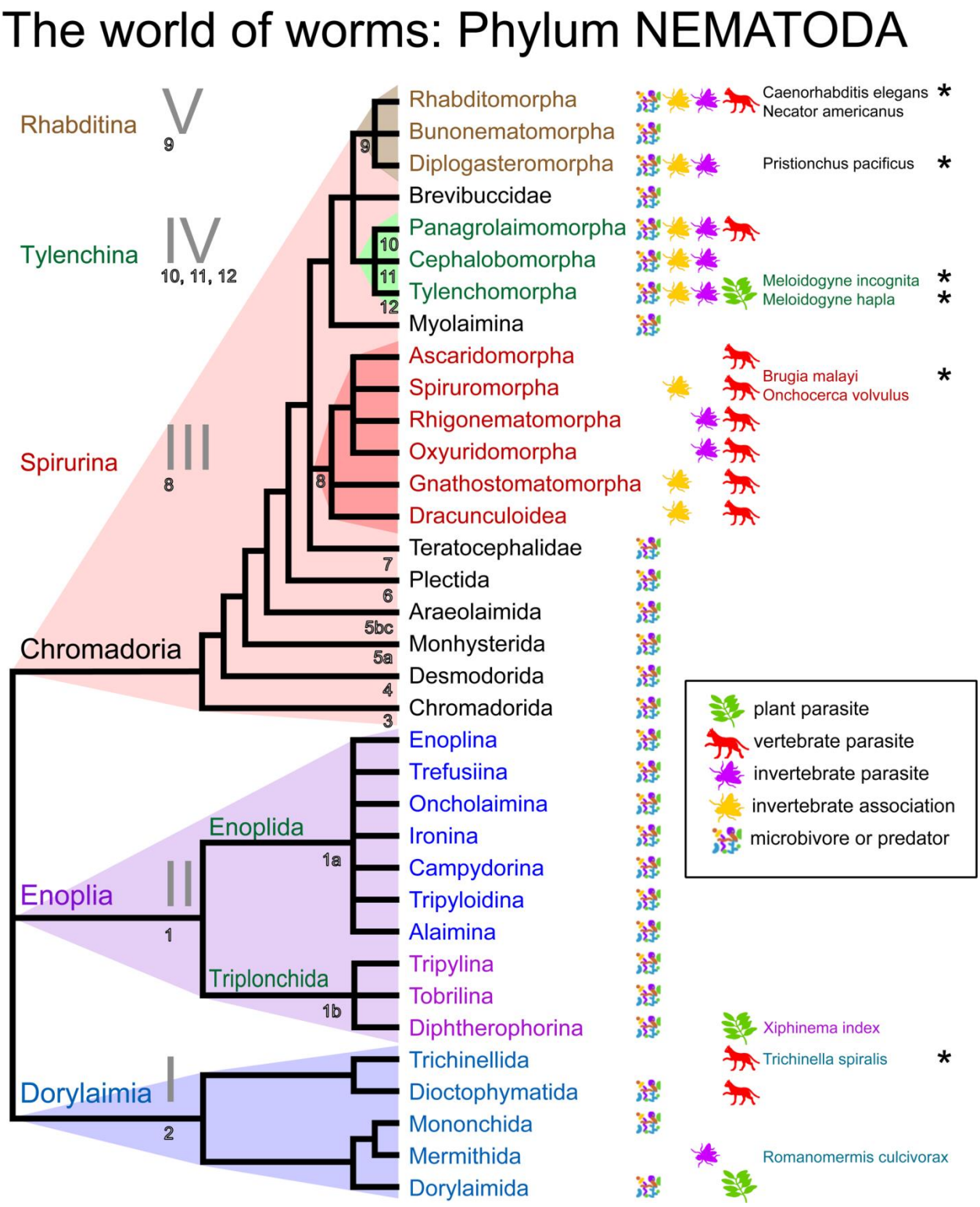


Figure 1. Nematode phylogeny based on SSU rRNA sequences. From Blaxter 2011.

Nematode surface antigen switching. Our laboratory developed two monoclonal antibodies, M37 and M38, that are specific for the surface of *C. elegans* L1 larvae (Hemmer et al 1991). In immunofluorescence staining experiments, M37 or M38 only stain the surface of L1 larvae, not the other larval stages or adults. Using these antibodies in immunofluorescence staining of live worms, a gene known as *srf-6* has been found to determine the timing of expression of antigens on the surface of the nematode (Hemmer et al 1991, Grenache et al 1996). *Srf-6* mutants were identified by a mutant phenotype in which all larval stages of *C. elegans* display the M37 epitope that is normally only found in L1 for wild type C. elegans (Hemmer et al 1991). The *srf-6* mutation was mapped to chromosome II. Display of the M37 epitope on the later larval stages is also induced in wild type worms grown on plates containing an extract of spent nematode liquid culture medium (Grenache et al 1996), suggesting that environmental cues can modulate surface antigen changes. Consistent with this interpretation, this induced larval display (ILD) requires the normal activity of chemosensory neurons (Olsen et al 2007).

In order to identify the DNA sequence corresponding to *srf-6*, whole genome sequencing of three different *srf-6* mutants was performed. Each of the three *srf-6* mutants was found to contain a different single base change in the coding sequence of previously identified gene *nsy-1* (Politz, Van Sciver, Rush, and Antoshechin unpublished results). Two bona fide *nsy-1* mutants showed expression of the M37 epitope at later larval stages, which is consistent with *nsy-1* and *srf-6* being the same gene (Bandursky, Isenhart, and Politz unpublished results).

NSY-1 is a *C. elegans* MAP kinase kinase kinase (MAPKKK), which is part of a mitogen activated protein kinase (MAPK) signal transduction pathway. In *C. elegans*, NSY-1 kinase activity is activated by calcium signaling by the upstream calmodulin-dependent protein kinase II encoded by the *unc-43* gene (Tanaka-Hino et al, 2002). *Nsy-1* has a role in determining whether the AWC olfactory neurons will be asymmetrical. In wild type *C.* elegans, one AWC neuron expresses the olfactory receptor *str*-2, and the other one does not. This has been detected by the asymmetric expression of *pstr-2:GFP* in only one AWC (Wes and Bargmann 2001). Among other genes, *nsy-1* activity promotes AWC being off and which AWC neuron of the two being off appears to be determined randomly (Troemel et al, 1999). Asymmetry is common in the development of *C. elegans,* occurring in other parts of the central nervous system as well as the AWC olfactory neurons. It has also been observed in *nsy-1* loss of function mutants, in which both AWC neurons express *str-2* (Troemel et al 1999, Wes and Bargmann 2001).

The human protein most similar in sequence to NSY-1 is ASK-1 (Apoptosis Signaling Kinase 1). ASK-1 is part of the human stress response to environmental factors such as osmotic

shock, oxidative conditions, and inflammation. Under normal cellular conditions, ASK-1 activity is inhibited by binding a dimer of thioredoxin at its N terminal domain. During an oxidative stress response, sulfhydryl groups of thioredoxin are oxidized, thereby releasing thioredoxin from ASK-1 (Hayakawa et al 2006, Figure 2). Upon activation, ASK-1 is used for cell death, immune regulation, cell differentiation, and cytokine responses (Hayakawa et al 2006).

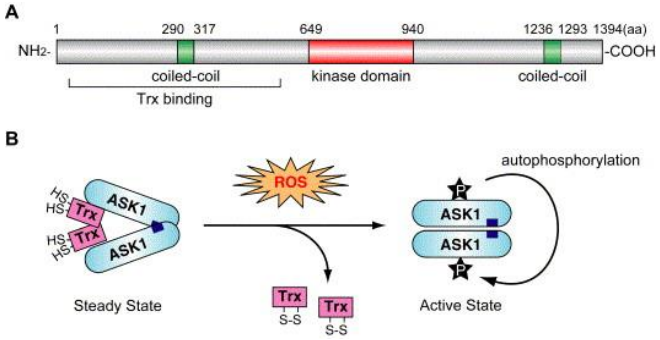


Figure 2: 2A) The coil-coiled domains and kinase domain in ASK-1. 2B) A basic representation of the thioredoxin reaction that regulates ASK-1 activity (from Hayakawa et al 2006).

The structure of ASK-1 includes three distinct domains (Figure 2). The N terminal domain contains the binding site for thioredoxin. The central domain contains the kinase catalytic domain, and the C terminal domain contains the dimerization site for the active ASK-1 dimer (Hayakawa et al 2006). In addition, both the N terminal and C terminal domains contain coiled coil protein association motifs. Most protein kinases are much smaller proteins than ASK-1, which may point to its importance in regulating stress responses.

Interestingly, the *C. elegans* NSY-1 protein does not seem to be regulated by thioredoxin, but by calcium signaling. Thus, one of the motivations of the present project was to look for sequence similarities in the non-catalytic domains of NSY-1-like proteins of other nematode species that might reflect similarities in regulation of activity.

Possible applications of this project. If parasitic nematodes have a gene similar to *nsy-1*, inhibiting this gene activity could turn off the changes in surface composition that occur during larval development. Since NSY-1 in *C. elegans* is part of the chemosensory pathway controlling the differentiation of AWC olfactory neurons (Sagasti et al, 2001) the inhibition of this enzyme could also lead to chemosensory interference in parasitic nematodes. In addition, preventing the switching of surface antigens of these parasites could allow the host immune system to better recognize and defend against the parasite. Eliminating parasites at early developmental stages might also avoid the pathogenesis that results from presence of parasitic worms in the body. Whether it is the breaking down of the lymphatic system (CDC, 2015) or stunted growth both physically and mentally (Bethony et al, 2006), the removal of these worms before reaching adulthood by the immune system could reduce these symptoms of worm infection.

To study how the *nsy-1* sequence is represented in parasitic nematodes, the program Geneious was used to search for similar sequences using the Basic Local Alignment Search Tool (BLAST). Once found, the sequences from parasitic nematodes as well as nematodes from the Caenorhabditis genus were aligned together to compare and search for consistent patterns within all the sequences. The final product was then seeing how these sequences placed the species in relation to *C. elegans* through the use of a phylogeny tree. By seeing how the different groups of parasitic nematodes are related to our model organism we can infer the species that are likely to be affected by methods that successfully interfere with the *nsy-1* gene in *C. elegans.*

Materials and Methods

The first part of this experiment used the sequence of *C. elegans nsy-1* to search for similar sequences in other parasitic and nonparasitic nematode species. Although this could have been done using the NCBI BLAST (Basic Local Alignment Search Tool) algorithm, the bioinformatics software platform Geneious was chosen because it integrates many different tools in a transparent, seamless environment. This allowed us to continue the analysis of similarity by alignment of sequences and phylogenetic tree-making without converting data at each step into appropriate formats. Geneious accessed files from the National Center for Biotechnology Information (NCBI) as well as other databases that keep protein sequence files from multiple publications. These databases included European Molecular Biology Laboratory (EMBL), DNA Databank of Japan (DDBJ), Protein Database (PDB), SwissProt, Protein Information Resources (PIR), and Protein Research Foundation (PRF) so both amino acid sequences and translated DNA sequences could be part of the final BLAST results (Geneious 7, http://www.geneious.com, Kearse et al., 2012). This way all possible sequences could be looked at to find the best possible alignments. It turned out that all of the sequences we included were from databases in NCBI.

We used the BLAST search integrated in Geneious, with the *C. elegans nsy-1* sequence as query. When a protein or DNA sequence is specified as query, this algorithm will go through databases that are based online such as NCBI. The results are sequences ordered in statistical significance to the query based on the similarity between the query and the results. As mentioned, the databases searched can contain nucleotide or protein sequences, and either type of database can be used regardless of the type of genetic material used as the query. For this experiment the *nsy-1* protein sequence was used as the query in a BLASTp search against multiple protein databases. All the sequences were found through the use of conceptually translated nucleic acid sequences such as cDNA and mtDNA (NCBI). The sources for the BLAST results were from multiple resources focused on sequencing proteins from the genomes of each species of nematode found in the results. There was a mix of sequence submission

sources that consisted of Consortiums as well as sequences being submitted that were part of research for further studying the nematodes by the lab publishing these results. It was also possible to specify the phylum *Nematoda* as the subject database. This narrowed the search to make sure all nematode proteins filed in NCBI were compared to *nsy-1* of *C. elegans*.

The choices of sequences to be aligned and then part of the final phylogeny tree were done first by organizing the BLAST results by increasing E value. The E value represents the probability that the similarity of two sequences found by BLAST occurs by chance, and it was calculated based on the length of the sequence being compared to the query and the number of amino acids or bases the query and sequence have in common (Karlin and Altschul 1990, Dembo et al, 1994).

Once the BLAST results were obtained, they were inspected. Those protein sequences closest in length to the *C. elegans nsy-1* sequence for each domain of the *nsy-1* protein sequence (N terminal, catalytic site, and C terminal) were selected to represent that species in the alignment and phylogeny tree. The goal was to narrow down the final sequences used to only include high-scoring segment pairs (HSPs), segments of equal or close to equal length to the original protein query for each domain (Sellers P 1984 and Gumbel E 1958). Some species, such as *Caenorhabditis briggsae*, had multiple BLAST results with an E value close to zero and had high percent similarity to *C. elegans* sequence. The sequence with a length closest to the length of *nsy-1* was chosen from the options even if that sequence had a smaller percent identity. The reason for this selection process is that sequences with smaller lengths later caused a misalignment with the other species based on the short length of one BLAST result. This would lead to conflicting results in the phylogeny trees; e.g., other closely related species in the genus *Caenorhabditis* were not within the same branch of the phylogeny tree as *C. elegans*.

In addition to an E value as close as possible to zero for each species, length similar to the *nsy-1* sequence and a relatively high similarity of no less than 50%, a requirement was added that the protein sequences from the BLAST results for each domain had to be from the same protein for each species represented in the phylogeny tree. The highest region of similarity for all species was in the catalytic domain. Therefore the proteins with highest similarity in the catalytic domain were also the proteins that were chosen in the N terminal and C terminal domains. However, if those proteins had high E values or were short in length for the N terminal or C terminal domains, they were not included in the alignment or phylogeny tree of those sections.

The species chosen according to these criteria were then aligned to locate regions of high similarity for each domain of the *nsy-1* sequence. This alignment was done using the program ClustalW. The program was used rather than the other alignment programs offered by

Geneious because ClustalW is a simple algorithm that is widely used for alignment. Also, the alternative alignment program (Geneious alignment) offers more variables that can affect the results of the alignment and with this experiment we wanted to keep the variables to a minimum. The reason for this is because when referencing these results in the future the conditions need to be easily replicated (Geneious 7, http://www.geneious.com, Kearse et al., 2012). ClustalW uses a BLOSUM62 matrix (Blocks substitution matrix) to align protein sequences through the numerical values given by how divergent the input sequences are. The BLOSUM62 matrix determines alignment of the sequence based on the amount of shared identity for each amino acid of the sequences being aligned and the number of amino acids that are substituted or differ between the compared sequences (Henikoff J and Henikoff S, 1992).

The phylogeny tree for the N terminal, C terminal, and catalytic domains of *nsy-1*-like protein sequences found in the BLAST search was constructed using the neighbor-joining method as well as the Jukes and Cantor distance model used by the Geneious tree-builder (Geneious 7, http://www.geneious.com, Kearse et al., 2012). The neighbor-joining method views the genetic material being compared and orders the species based on the law of parsimony, or through the tree which gives the smallest number of evolutionary changes. In order to find the number of evolutionary changes, the Jukes and Cantor distance model was used to determine the frequency of nucleotide substitution (Van de Peer, Y. 2009). This model along with the Neighbor-joining method provided the algorithms that resulted in the final phylogeny tree for each section of *nsy-1* protein sequence. The designated outgroup was *Homo sapiens* ASK1 because it is the human protein sequence most closely related to *C. elegans nsy-1*.

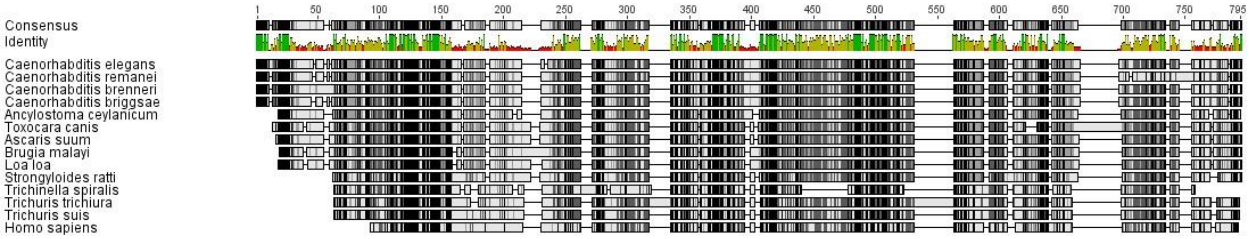
Upon analysis of the *nsy-1* sequence for motifs a coiled-coil predictor was used to see if the sequence contained any coil-coil motifs, because ASK-1 contains coiled-coil motifs as part of both the C terminal and N terminal domains of similarity in that certain area for all species included in the alignment. The coiled-coil predictor used to find this particular domain utilized various protein prediction programs to give a probability as to the type of protein structure being coded for in the sequences given. This particular domain being searched for was coiled-coil. The programs database included protein sequences of known globular proteins, randomly generated sequences, and all sequences in GenBank. (Lupas et al et al, 1991).

Results

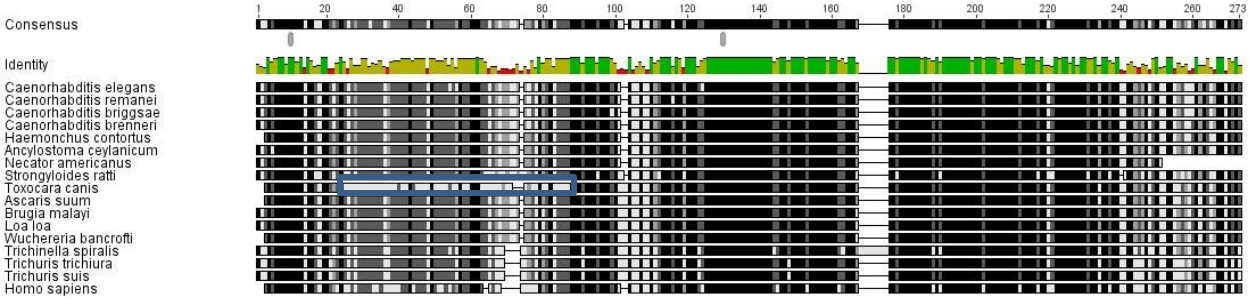
The analysis of how each nematode species found from the BLAST search was related to our query was done using the alignments of their protein sequences to NSY-1 of *C. elegans.* The *Homo sapiens* (*H. sapiens*) sequence was also included as an outgroup (Figure 4). From these alignments phylogeny trees were made for each section of *nsy-1's* sequence to compare

the locations of each species based on the different sequence segments (N terminal domain, catalytic domain, and C terminal domain) (Figure 3).

A.



B.



C.

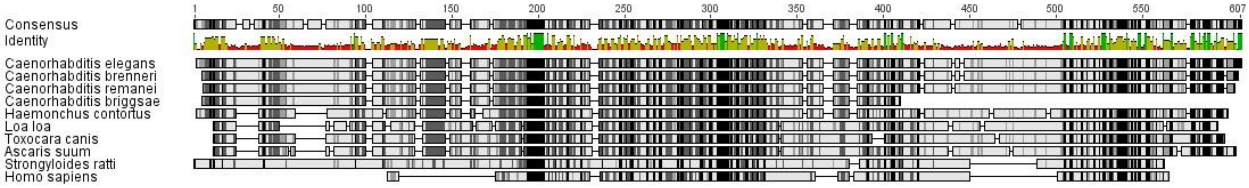


Figure 3: 3A) an alignment of NSY-1's N terminal domain with species collected from a BLAST search using NCBI's database of sequences through Geneious. 3B) an alignment of NSY-1's catalytic domain with species that were found using a BLAST search with the program Geneious connected to NCBI's database. A noticeable difference in the general pattern of alignment is outlined by a blue box; this occurs in the sequence of *Toxocara canis*. 3C) An alignment of NSY-1's C terminal sequence with matches from species found using a BLAST search with the program Geneious connected to NCBI's database. Sequence similarities among the given sequences are shown in two ways. First, a scale of light to dark shading is assigned based on how many matches exist amongst each amino acid in the sequences of the alignment. If the alignment shows a high similarity of one species to the others, the amino

acids are colored black for that sequence region. Nonidentical amino acids that represent conservative substitutions amongst the species are shown in grey. Grey is also used to indicate a section in which some but not all amino acids for that section match, with a gradation of shading from dark to light. Second, similarity is indicated by shown by the colored graph between the consensus sequence and the first species in the alignment list. Each column of the amino acid sequences once aligned is graded on the percentage of similarity and that grade is represented by the color of the graph. If green all amino acids present in the column have high percent of similarity at 60% or more compared to the nsy-1 sequence of C. elegans. A brown color shows 40-60% similarity for the column of amino acids being compared and red shows less than 40% similarity. (Geneious version 7, http://www.geneious.com, **Kearse et al., 2012)**

To further analyze how each species is related not only to *C. elegans* but to the other species present, a table that showed the calculated similarity between each species was created (Table 1). The numerical values given helped visualize the relationship between species and also gave some surprising results, for example in the N terminal domain, *H. sapiens* sequence is more closely related to *C. elegans* than *T. spiralis* is related to *C. elegans*. These results provided a way to understand how the alignments in Figure 4 became the phylogenetic trees given in Figure 4.

A.

	C. elegans	C. remanei	C. brenneri	C. briggsae	A. cyclostoma	T. canis	A. suum	B. malayi	L. loa	S. ratti	T. spiralis	T. trichuria	T. suis	H. sapiens
C. elegans		86.095	91.904	90.06	70.015	60.231	63.238	61.151	61.854	54.428	33.731	39.728	42.997	36.661
C. remanei	86.095		86.873	84.718	64.34	54.674	57.355	54.888	55.373	49.289	30.551	35.905	38.818	33.052
C. brenneri	91.904	86.873		91.291	69.029	59.424	62.538	60.759	61.457	53.945	33.22	38.973	42.182	37.177
C. briggsae	90.06	84.718	91.291		70.17	59.078	62.33	60.398	60.942	54.267	33.39	39.124	42.345	37.522
A. cyclostoma	70.015	64.34	69.029	70.17		63.982	67.385	65.403	66.462	56.792	34.542	39.144	42.409	35.602
T. canis	60.231	54.674	59.424	59.078	63.982		90.469	79.423	79.853	55.59	33.657	38.873	41.77	34.804
A. suum	63.238	57.355	62.538	62.33	67.385	90.469		84.976	85.626	60	36.068	40.819	44.026	37.024
B. malayi	61.151	54.888	60.759	60.398	65.403	79.423	84.976		95.686	57.382	34.974	39.947	43.049	37.102
L. loa	61.854	55.373	61.457	60.942	66.462	79.853	85.626	95.686		58.101	35.959	40.881	44.098	37.716
S. ratti	54.428	49.289	53.945	54.267	56.792	55.59	60	57.382	58.101		33.673	38.847	42.062	35.406
T. spiralis	33.731	30.551	33.22	33.39	34.542	33.657	36.068	34.974	35.959	33.673		48.445	52.189	30.642
T. trichuria	39.728	35.905	38.973	39.124	39.144	38.873	40.819	39.947	40.881	38.847	48.445		90.031	37.175
T. suis	42.997	38.818	42.182	42.345	42.409	41.77	44.026	43.049	44.098	42.062	52.189	90.031		40.317
H. sapiens	36.661	33.052	37.177	37.522	35.602	34.804	37.024	37.102	37.716	35.406	30.642	37.175	40.317	

B.

	C. elegans	C. briggsae	C. remanei	C. brenneri	N. americanus	A. ceylanicum	H. contortus	S. ratti	T. canis	A. suum	B. malayi	L. loa	W. bancrofti	T. spiralis	T. trichuri	T. suis	H. sapiens
C. elegans		92.748	93.13	91.985	87.917	86.641	87.692	72.348	61.45	77.863	78.048	77.273	78.244	66.3	70.189	70.189	65.134
C. briggsae	92.748		92.366	90.84	87.917	85.878	86.538	73.106	63.359	80.153	78.426	77.652	78.626	67.033	70.566	70.566	63.602
C. remanei	93.13	92.366		93.13	89.167	87.786	88.846	73.485	63.74	80.916	79.941	79.545	80.534	67.033	70.943	70.943	65.134
C. brenneri	91.985	90.84	93.13		90	88.55	89.231	74.242	61.069	78.626	76.911	76.136	77.481	65.934	69.434	69.434	64.368
N. americanus	87.917	87.917	89.167	90		98.333	99.16	75.207	66.667	84.167	82.663	81.818	82.917	67.331	72.428	72.428	66.109
A. ceylanicum	86.641	85.878	87.786	88.55	98.333		98.462	73.485	65.649	81.298	81.078	80.303	80.916	66.667	71.698	71.698	65.134
H. contortus	87.692	86.538	88.846	89.231	99.16	98.462		73.282	66.031	81.679	81.315	80.534	81.298	67.159	71.483	71.483	64.751
S. ratti	72.348	73.106	73.485	74.242	75.207	73.485	73.282		57.034	73.384	72.093	71.321	72.624	66.3	67.547	67.925	61.832
T. canis	61.45	63.359	63.74	61.069	66.667	65.649	66.031	57.034		77.481	73.3	73.664	73.664	60.223	62.452	62.452	59.77
A. suum	77.863	80.153	80.916	78.626	84.167	81.298	81.679	73.384	77.481		93.529	93.511	93.893	74.17	76.426	76.426	67.681
B. malayi	78.048	78.426	79.941	76.911	82.663	81.078	81.315	72.093	73.3	93.529		98.881	99.636	73.643	75.111	75.111	67.698
L. loa	77.273	77.652	79.545	76.136	81.818	80.303	80.534	71.321	73.664	93.511	98.881		99.237	73.993	75.472	75.472	68.061
W. bancrofti	78.244	78.626	80.534	77.481	82.917	80.916	81.298	72.624	73.664	93.893	99.636	99.237		74.17	74.905	74.905	67.681
T. spiralis	66.3	67.033	67.033	65.934	67.331	66.667	67.159	66.3	60.223	74.17	73.643	73.993	74.17		82.156	82.156	67.041
T. trichuria	70.189	70.566	70.943	69.434	72.428	71.698	71.483	67.547	62.452	76.426	75.111	75.472	74.905	82.156		98.851	67.181
T. suis	70.189	70.566	70.943	69.434	72.428	71.698	71.483	67.925	62.452	76.426	75.111	75.472	74.905	82.156	98.851		67.181
H. sapiens	65.134	63.602	65.134	64.368	66.109	65.134	64.751	61.832	59.77	67.681	67.698	68.061	67.681	67.041	67.181	67.181	

C.

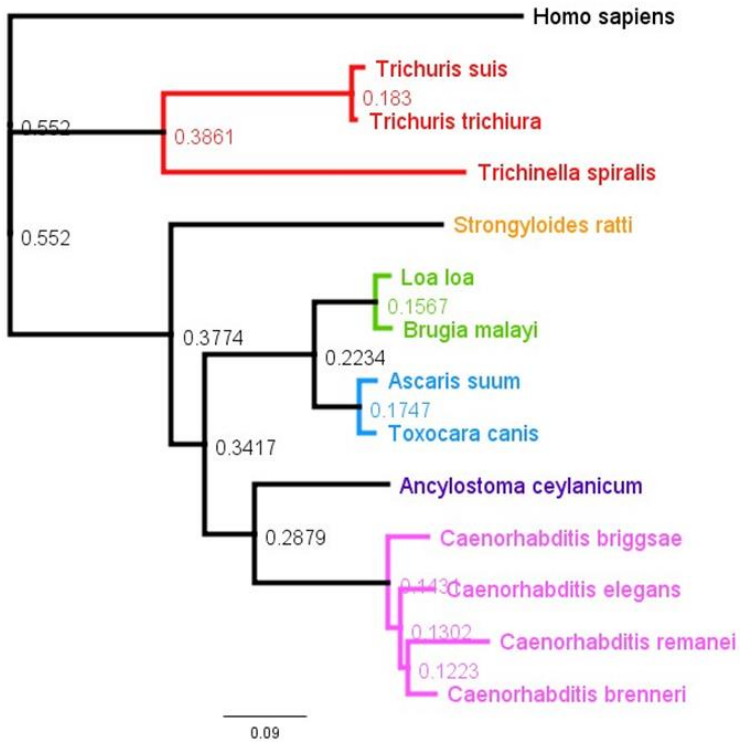
	C. elegans	C. brenneri	C. remanei	C. briggsae	H. contortus	L. loa	T. canis	A. suum	S. ratti	H. sapiens
C. elegans		93.134	92.756	91.534	51.049	35.889	38.302	38.87	17.594	20.554
C. brenneri	93.134		93.451	92.063	50.791	35.54	37.955	38.185	18.051	20.554
C. remanei	92.756	93.451		91.512	51.761	36.411	38.475	39.108	18.264	20.092
C. briggsae	91.534	92.063	91.512		53.786	39.487	41.645	42.674	16.958	23.105
H. contortus	51.049	50.791	51.761	53.786		37.044	40.109	39.964	16.216	19.672
L. loa	35.889	35.54	36.411	39.487	37.044		62.662	64.739	18.182	24.379
T. canis	38.302	37.955	38.475	41.645	40.109	62.662		84.926	18	21.622
A. suum	38.87	38.185	39.108	42.674	39.964	64.739	84.926		18.182	22.748
S. ratti	17.594	18.051	18.264	16.958	16.216	18.182	18	18.182		15.854
H. sapiens	20.554	20.554	20.092	23.105	19.672	24.379	21.622	22.748	15.854	

Table 1: The percent BLAST similarity between each pair of species. The shading indicates similarity; the more similar two species are the darker the cell showing the percent similarity. The species are from the BLAST search in which *C. elegans* NSY-1 was the query. 1A) the sequences being compared are from the N terminal domain. 1B) the sequences being compared are from the catalytic domain BLAST results. 1C) the species being compared are from the C terminal domain BLAST results. (Geneious version 7, http://www.geneious.com, Kearse et al., 2012)

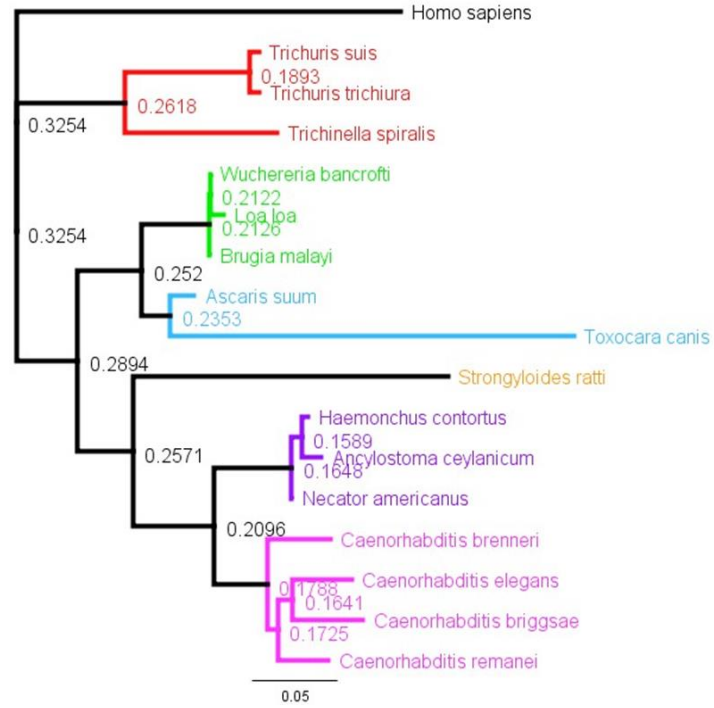
This degree of high similarity between species in the catalytic domain compared to the species being compared in the N and C terminal domains can be seen in Table 1B by looking at the percentage of similarity as each species is compared to the others. In Table 1B there many black cells compared to the N terminal and C terminal with no species being less than 50% similar. While there are species with high similarity in the N terminal and C terminal comparisons, the species that are least similar can reach as low as 18% similarity (Figure 3C, and Table 1C). This reinforces the results seen in Figures 3A-3C where the catalytic domain shows the highest degree of similarity between species.

The high percentage of similarity as well as the long stretches of matching sequences seen in Table 1 and Figure 3, respectively, is mirrored in phylogenetic trees, as the same species are shown to be closely related (Figure 4A-4C). A specific example can be seen with the species *B. malayi* and *L. loa*. These two species have 95% similarity according to the N terminal and catalytic domain comparison (Table 1A and Table 1B). In addition by looking at Figure 4A they have five common ancestors. A common ancestor between a group of species is the point where the phylogenetic tree breaks into two separate branches. The common ancestor for all species being compared in the tree is the starting point of the tree, and if you follow a path of a tree to a certain species, the common ancestors it shares with another species being compared is the number of nodes different species, not being compared, branch off from until the two species being compared form their own separate branches. In the case of *L. loa* and *B. malayi* there are five nodes until they are placed on two separate branches (Figure 4A). The high percentage of similarity between species leaves the C terminal section with the fewest sections of high similarity between the three domains (Table 1C). By looking at Table 1C the species most closely related were within the Caenorhabditis genus and even the species that are second in similarity within the C terminal domain show the percentage of similarity to be in the low 80s. After that the highest percent of similarity is in the 50s and this shows just how dissimilar the sequences in the C terminal domain are compared to the sequences being compared in the N terminal and catalytic domain (Tables 1A-1C). At the same time the N terminal domain has the most sequences that are not of length similar to *C. elegans* NSY-1 in *C. elegans* (Figure 3A). That fact notwithstanding, the percent similarity between species in the C terminal is the lowest of the three protein domains.

A.


Homo sapiens
Trichuris suis
0.183
Trichuris trichiura
0.552
0.3861
Trichinella spiralis
0.552
Strongyloides ratti
Loa loa
0.1567
Brugia malayi
0.3774
0.2234
Ascaris suum
0.1747
Toxocara canis
0.3417
Ancylostoma ceylanicum
0.2879
Caenorhabditis briggsae
Caenorhabditis elegans
0.1302
Caenorhabditis remanei
0.1223
Caenorhabditis brenneri
0.09

B.


Homo sapiens
Trichuris suis
0.1893
Trichuris trichiura
0.2618
0.3254
Trichinella spiralis
Wuchereria bancrofti
0.2122
Loa loa
0.2126
0.3254
Brugia malayi
0.252
Ascaris suum
0.2353
Toxocara canis
0.2894
Strongyloides ratti
Haemonchus contortus
0.1589
Ancylostoma ceylanicum
0.2571
0.1648
Necator americanus
0.2096
Caenorhabditis brenneri
Caenorhabditis elegans
0.1641
Caenorhabditis briggsae
0.1725
Caenorhabditis remanei
0.05

C.

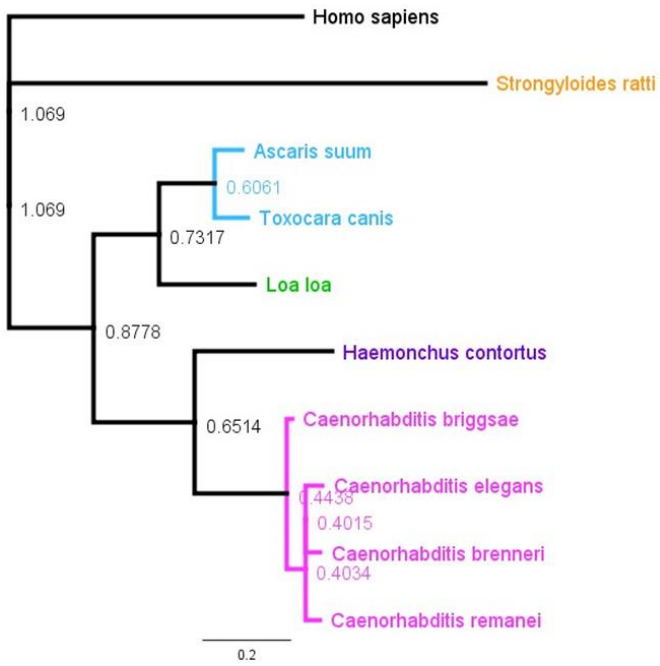


Figure 4: Phylogenetic trees created using alignments of the NSY-1 sequence from the BLAST results, in which NSY-1 of *C. elegans* was used as the query as shown in Figures 3A-3C. Geneious was used to find the algorithms and compile the data to create these phylogeny trees. The data used to make the trees came from the multiple alignment performed on the BLAST results. 4A) a phylogenetic tree created using the alignments of the NSY-1 N terminal sequence; 4B) a phylogenetic tree made from the catalytic domain sequence; 4C) a phylogenetic tree made from the C terminal sequence. (Geneious version 7, http://www.geneious.com, Kearse et al., 2012)

Position of species in the Phylogeny of Nematodes

Figure 4 show the N terminal domain phylogeny, catalytic domain phylogeny, and the C terminal domain phylogeny. All of the trees show that, as expected, the non-parasitic free living soil nematodes are most closely related to *C. elegans*. By grouping the different species by common taxa (Figure 4A-4C and Figure 5) we can see that the suborder Strongylida (Uniprot, 2015) makes up the nematodes most closely related to *C. elegans* that contain parasitic nematodes. For the N terminal domain, this order is represented by the branch containing *A. ceylanicum*. The suborder that follows the Strongylida in similarity is Ascaridida, making up the branch *A. suum* and *T. canis*, and this order shares a common ancestor with the Spirurida order consisting of *L. loa* and *B. malayi* (Uniprot, 2015 and Figure 5). *S. ratti* is in order Rhabditida (Figure 5), and this species shares a common order with all the species mentioned previously, (Uniprot, 2015). *T. spiralis*, *T. suis*, and *T. trichura* and these species makes up the

Trichocephalida order (Uniprot, 2015) (Figure 5) that is the only representative in our data that represents class Enoplea. All the other species found are in Class Chromodorea.

Phylum	Class	Subclass	Order	Suborder	Super Family	Family	Subfamily	Genus	Species
Nematoda	Chromadorea		Rhabditida	Rhabditoidea	Rhabditidae	Peloderinae		Caenorhabditis	C. elegans
Nematoda	Chromadorea		Rhabditida	Rhabditoidea	Rhabditidae	Peloderinae		Caenorhabditis	C. briggsae
Nematoda	Chromadorea		Rhabditida	Rhabditoidea	Rhabditidae	Peloderinae		Caenorhabditis	C. brunnei
Nematoda	Chromadorea		Rhabditida	Rhabditoidea	Rhabditidae	Peloderinae		Caenorhabditis	C. remanei
Nematoda	Chromadorea		Rhabditida	Strongylida	Ancylostomatoidea	Ancylostomatidae	Ancylostomatinae	Ancylostoma	A. ceylanicum
Nematoda	Chromadorea		Rhabditida	Strongylida	Ancylostomatoidea	Ancylostomatidae	Bunostominae	Necator	N. americanus
Nematoda	Chromadorea		Rhabditida	Strongylida	Trichostongyloidae	Haemonchidae	Haemonchinae	Haemonchus	H. contortus
Nematoda	Chromadorea		Rhabditida		Panagrolaimoidea	Strongyloididae		Strongyloides	S. ratti
Nematoda	Chromadorea		Rhabditida	Ascaridida	Ascaridoidea	Toxocaridae		Toxocara	T.canis
Nematoda	Chromadorea		Rhabditida	Ascaridida	Ascaridoidea	Ascarididae		Ascaris	A. suum
Nematoda	Chromadorea		Spirurida	Filarioidea		Onchocercidae		Wuchereria	W. bancrofti
Nematoda	Chromadorea		Spirurida	Filarioidea		Onchocercidae		Brugia	B. malayi
Nematoda	Chromadorea		Spirurida	Filarioidea		Onchocercidae		Loa	L.loa
Nematoda	Enoplea	Dorylaimia	Trichocephalida			Trichnellidae		Trichinella	T. spiralis
Nematoda	Enoplea	Dorylaimia	Trichocephalida			Trichuridae		Trichuris	T. suis
Nematoda	Enoplea	Dorylaimia	Trichocephalida			Trichuridae		Trichuris	T. trichuris

Figure 5: Phylogenetic organization of nematode taxa relative to the taxon of *C. elegans.* The taxa that the species have in common with *C. elegans* are in red. The species group that consists of *Necator americanus, Haemonchus contortus,* and *Ancylostoma ceylanicum* falls under the suborder Strongylida and that suborder is shown in orange. The species group that includes the filarial nematodes, *Wuchereria bancrofti*, *Brugia malayi,* and *Loa loa* is the taxa order Spirurida and is colored yellow. The species that are included in the suborder Ascaridida are *T. canis* and *A. suum.* The species grouped into order Trichocephalida are *Trichinella spiralis, Trichuris suis,* and *Trichuris trichuris* and this is colored in blue. *Strongyloides ratti* has no common taxa with any of the parasitic species below the level of order, and the highest taxa that this species has in common with *C. elegans* is the order Rhabditida. The taxonomy of the species was found courtesy of (Uniprot, 2015).

N terminal description

The N terminal alignment (Figure 3A) shows areas where gaps in the sequences are uniform among some of the species. Such sections are amino acids 530-560 as well as 640-690. This may be a result from the insertions and deletions within sequences that deviate more from *nsy*-1's sequence. These deletions and insertions show which sequences are still matching consistently with our query and adds to the overall picture of the placement of each species in relation to *C. elegans.* The starting points of the sequences are not uniform among all of the species. The species in genus Caenorhabditis match the *C. elegans* query. However, the *Ancyclostoma ceylanicum*, *Loa loa*, *Brugia malayi*, *and Ascaris suum* start after the first 50 amino acids of the query, and the *Trichuris* species, *Trichinella spiralis*, and *Strongyloides ratti* sequences start at amino acids of the query. *Homo sapiens* sequence starts just before the 100 amino acid mark of the query. From comparing the alignments of the sequences to their placement on the phylogenetic tree, a sequence may have high amounts of similarity to the

query but if it is shorter in sequence length the species will be considered less closely related than a species with the same regions of similarity but of length closer to that of *nsy-1*. For example *A. ceylanicum*'s sequence has the same areas of high similarity as the species that are part of the Caenorhabditis genus (Figure 3A), but its sequence is shorter and therefore is less closely related to *C. elegans* than the non-parasitic free living nematode species that are part of the phylogenetic tree (Figure 4A).

Catalytic domain description

The catalytic domain for *nsy-1* has a greater number of species containing proteins with high similarity to the *C. elegans* query than the N terminal and C terminal domains. Unlike the two terminal domains of *nsy-1,* the catalytic site shows strong similarity with *Necator americanus* (*N. americanus*) with a sequence length of approximately the same length as *nsy-1* in *C. elegans*. Also Figure 3B shows that the catalytic domain of *nsy-1* has very few areas of low similarity between species; the main gap in the sequence similarity is for the amino acid segment 165-175, and this is caused by a sequence inserted in the distantly related species *Trichinella spiralis* Most species in the alignment follow the same patterns of similarity and gaps but *Toxocara canis* appears to have large gaps in its sequence that do not follow the overall pattern in the first 100 amino acids (this is outlined in blue in Figure 3B).

The placement of different common taxa groups in relation to *C. elegans* for the catalytic domain, seen in Figure 3B, once again shows that after the non-parasitic free-living nematodes that are in the same genus as *C. elegans*, the suborder Strongylida, represented by *A. ceylanicum, N. americanus*, and *H. contortus* (Figure 5) is most closely related to *C. elegans*. This proximity is the same as seen in Figure 4A with the phylogenetic tree using the N terminal domain, although in the latter, the sole representative of Strongylida is *A. ceylanicum*. The location of *S. ratti,* which is part of the Rhabditida order, like the other aforementioned species, is different though, as this species sequence for the catalytic domain is closer in relation to *C. elegans* than Ascaridida, a suborder that is represented by *T. canis* and *A. suum,* and Spirurida, an order represented by the species *W. bancrofti*, *L. loa*, and *B. malayi* (Figure 5). Both Ascaridida and Spirurida are more closely related to *C. elegans* than *S. ratti* when comparing the N terminal in Figure 4A but what remains the same is the location of the Spirurida order, seen less closely related to *C. elegans* than the Ascaridida order when looking at the phylogeny tree in Figure 4. In addition, Trichocephalida, represented by species *T. suis*, *T. spiralis*, and *T. trichuria* (Table 1) is least related to *C. elegans*, being closest to the outgroup, *H. sapiens*.

C terminal description

Figure 3C shows where the C terminal domain has high similarity between the proteins for each species being compared to NSY-1of *C. elegans*. When just looking at how similar the sequences are, high similarity does exist for amino acids 175 to approximately 330 but there is no other portion of the alignment that has a high similarity between all of the species present in the alignment.

Looking at the phylogenetic tree of NSY-1's C terminal sequence (Figure 4C), the results appear to differ noticeably from Figure 4B. In Figure 4B, S. *ratti* was less closely related to the *Caenorhabditis* genus than were the species representing the suborder Stronglydida (Figure 5). In contrast, for the C terminal phylogeny shown in Figure 4C, *S. ratti* is the most distantly related of all the nematode species represented in relation to the *Caenorhabditis* species. This is more noticeable when comparing the C terminal sequence to the catalytic domain because if the Trichocephalida order was removed from Figure 4A for the N terminal domain's phylogenetic tree, then *S. ratti* would be in the same position in both figures. In addition the Spirurida order, represented in Figure 4C by *L. loa* (Figure 4), is more closely related to *C. elegans* than suborder Ascaridida, and this does not occur in Figures 4A or 4B based on the positioning of species representing the Spirurida order. What does remain constant for all phylogenetic trees though is the species representing the suborder Stronglydida are the most closely related parasitic nematodes to *C. elegans* (Figures 4A-4C).

A problem that occurs not only in Figure 4C but Figure 4A as well is that not all of the species represented in the catalytic site's phylogeny tree (Figure 4B) are represented in the N terminal and C terminal phylogeny trees. This can be difficult when comparing placement of species, but besides order Trichocephalida, all orders are represented in every phylogenetic tree. Trichocephalida is not seen in the C terminal tree (Figure 4C) but at the same time this means the BLAST results were not high enough in identity percentage or long enough in sequence length. This fact and that the placement of this order is the same for both Figure 4A and Figure 4B could suggest that the species that make up this order do not express proteins closely related to *C. elegans nsy*-1.

Mutations

The amino acid locations of known *nsy-1* mutations *ky400, ky542,* and *ky397* (Sagasti et al 2001) and *srf-6* mutations *yj13, yj41,* and *yj15* (S. Politz, personal communication) were compared in the species found from the BLAST results to see if the amino acids in other species were the same for these locations, and whether their location was in highly similar sequences. Mutations conferring the Srf-6 and Nsy-1 phenotypes were found in all sections of the *nsy-1* sequence (N terminal, catalytic, and C terminal domains). These mutations are shown in Figure 6.

A. Mutation *yj13*

```
Ce wt           NPDVLHPDTVSLMMLSY-----------------RDNQNYGGMIRLVDDLKRIPDCLKVV
Ce yj13         NPDVLHPNTVSLMMLSY-----------------RDNQNYGGMIRLVDDLKRIPDCLKVV

T. SUIS         DPDVLSVDTVVNMLLSY-----------------RDIQDYAAMVSLVEDIRSIPNNK-IY
T. TRICHIURA    DPDVLSVDTVVNMLLSYSATLFSSHLTNLPLISLRDIQDYAAMVSLVEDIRSIPNNK-IY
T. SPIRALIS     RGYCSQHAFVVQVQILFVQ---------------LDIQNYNAMVSLVEDVSSLPYDK-IH
C. REMANEI      NPDVLHPDTVSLMMFSY-----------------RDNQNYGGMIRLVDDLKRVTDCVKVV
C. BRENNERI     NPDVLHPDTVSLMMLSY-----------------RDNQNYGGMIQLVDDLKRVTDCVKVV
C. BRIGGSAE     NPDVLHPDTVSLMMLSY-----------------RDNQNYGGMIRLVDDLKRITDCVKVV
A. CEYLANICUM   NPDVLSVDTVHQLMISY-----------------RDNQNYNGMISLVEDLSRIEDCT-LI
L. LOA          NPDVLSVDTLYQFMLSY-----------------RDYQNYDAMISLYDDLSRIENCS-IV
B. MALAYI       NPDVLSVDTLYQFMLSY-----------------RDYQNYDAMISLYDDLSRIENCS-IV
A. SUUM         NPDVLSVDTLYQFMLSY-----------------RDNQNYDAMISLIDDLSRIENCS-IV
S. RATTI        HPDVLSVDTLHQYLLSY-----------------RDTENYIGMISLIDDLLKIGQHQ-IV
T. CANIS        NPDVLSVDTLYQFMLSY-----------------RDNQNYDAMISLIDDLSRIENCC-IV
H. SAPIENS      NTEVLTSDIIINLLLSY-----------------RDIQDYDAMVKLVETLEMLPTCD-LA
```

B. Mutation *ky400*

```
Ce wt         FMEATE----ADTDISCPRYPVLILELNKEFTPSYLTLN—NEEGTVILSHVLENSQQK
Ce ky400      FMEATE----ADTDISCPRYPVLILELN*EFTPSYLTLN--NEEGTVILSHVLENSQQK

T. SUIS       FVESTK----ADDLHCDLRFPVLSMEPSKAYTPSYVSVN--KDDKTIFLWHVTYSLPKS
T. TRICHIURA  FVESTK----ADDLHCDLRFPVLSMEPSKAYTPSYVSVN--KDDKTIFLWHVTYSLPKS
T.SPIRALIS    FVENTK----SDEICTDTRFPVLTMEPSKIYIPSYVLVN--EESKAISLWHVEDTNRK-
C. REMANEI    FMEATE----VDTDINCPRYPVLILELNKEFTPSYLTLN--NEEGTVILSHVLENSQQK
C. BRENNERI   FMEATE----ADTDINCPRYPVLILELNKEFTPSYLTLN—NEEGTVILSHVLENSQQKK
C. BRIGGSAE   FMEATE----TDTDINCPRYPVLILELTKEFTPSYLTLN—TEEGTVILSHVLETSQQKK
A. CEYLANIC   FMEAID----SETDVTCARFPVLIQELTKQFTPSYLTLN—VTERSMILSHVLENSQQKK
L. LOA        FMEAID----SEEEIVCGRFPVLIQEVTKQYTPSFLTLN--VGEGSIILSHVLESSQQK
B. MALAYI     FMEAID----SEEEIVCGRFPVLIQEVTKQYTPSFLTLN--VSEGSIILSHVLESSQQK
A. SUUM       FMEAID----SESEIVCARFPVLIQEVTKVYTPSFLTMN--LSEGSIILSHVLENSQQK
S. RATTI      FMEAIENTKPYENDIECLRVPVLIQEVNKQYTPSYLSIN--LKEGSVILFHVRENAPST
T. CANIS      FMEAID----SESEIVCARFP--------VYTPSFLTMN--LSEGSIILSHVLENSQQK
H. SAPIENS    IFEATN------EVTNGLRFPVLVIEPTKVYQPSYVSINNEAEERTVSLWHVSPTEMKQ
```

C. Mutation *ky542*

```
Ce wt           QLFLYVHENSDDFN---LLFPTKAH---------CKKAYDDMKSMADVADGNYQGRVLS
Ce ky542        QLFLYVHENSDDFN---LLFPTKAH---------CKKAYDDMKSMADVADGNY*GRVLS
```

```
T. SUIS          AVYLYVYLNSDDFM---LFFPTDSH---------RHRFLDLVDEITSNVDG---TKLLG
T. TRICHIURA     AVYLYVYLNSDDFM---LFFPTDSH---------RHRFLDLVDEITSNVDG---TKLLG
T. SPIRALIS      AVYLYVYLNSDDFM---LFFPSDNH---------RKR-----------------------
C. REMANEI       FMFMKIQMISTFFSQQKLIVKSKNSKFDSSSHFIFFRAFDDMKSMADVADGNYQGKVLS
C. BRENNERI      QLFLYVHENSDDFN---LLFPTKAH---------CKKAFDDMKSMADVADGNYQGKVLS
C. BRIGGSAE      QLFLYVHENSDDFN---LLFPTKAH---------CQKAFDDMKSMADVEDGNYQGKVLS
A. CEYLANICUM    SMFLYVHENSDDFN---LVFPSAAH---------CNKVINTLMEMTEADG--TQGKVLS
L. LOA           SMFLYVHENSDDFN---LVFPSAAH---------CNKVMGSLIEMINELDG-NAGRVLH
B. MALAYI        SMFLYVHENSDDFN---LVFPSAAH---------CNKVMGSLIEMINELDG-NAGRVLH
A. SUUM          SMFLYVHENSDDFN---LVFPSAAH---------CNKVMASLIEMTSELDG-NAGRVLH
S. RATTI         SMYLYVHENSDDFN---LVFPNAEI---------CTKIIAMISHLDCDVD-----KVLP
T. CANIS         SMFLYVHENSDDFN---LVFPSAAH---------CNKVMASLIEMTSELDG-NAGRVLH
H. SAPIENS       CCFLYVHDNSDDFQ---IYFSTEEQ---------CSRFFSLVKEMITNTAG—STVELEG
```

Figure 6: The aligned amino acid sequences of species with HSPs in the N terminal domain of *C. elegans* NSY-1. The mutant amino acid is colored red. 6A) Mutation *yj13* converts wild-type aspartic acid to mutant asparagine 6B) Mutation *ky400* converts a lysine in wild type *C. elegans* to a premature stop codon. 6C) Mutation *ky542* occurs at a glutamine amino acid in wild-type *C. elegans* protein sequence and also results in a premature stop codon.

The N terminal domain contained three separate locations where point mutations were induced, and when comparing these amino acids to the other species the uniform identity or lack thereof is noticed. Figure 6 shows that for mutation *yj13*, wild-type aspartic acid (D) is the amino acid that is induced to mutate in NSY-*1*'s sequence. In *yj13,* the aspartic acid is converted to asparagine. This is an interesting change with respect to protein function, because aspartic acid has a negatively charge side chain, and asparagine has a positive charge. An aspartic acid is consistently found at this position for almost all of the species being compared to *C. elegans.* The exception is for *T. spiralis* which has an alanine (A) in this location instead of aspartic acid. In addition to conservation of the amino acid modified by *yj13*, the entire sequence flanking the mutation site is highly similar in all of the species.

For mutation *ky400* every species except *T. canis* has the wild type amino acid lysine at the point where the mutation was induced (Figure 6B) The *ky*400 mutation creates a stop codon at this point. *T. canis* has a gap which deletes this position. In spite of this, the overall sequence surrounded *ky400* is similar in most of the species.

Figure 6C shows the position at which mutation *ky542* was induced; the wild type amino acid at this location is glutamic acid (Q) for the *Caenorhabditis sp.* and *A. ceylanicum*. The *C. elegans ky400* mutation creates a stop codon at this point. The rest of the species that do not have glutamic acid at this position either have a gap, alanine, or tyrosine. The species *T. trichuria, T. suis,* and *T. spiralis* all have gaps at this position while *A. ceylanicum, L. loa, B. malayi,* and *T. canis* have alanine at this mark. *H. sapiens* is the only species with tyrosine (T) at this location. This shows considerable variation in comparison to the amino acids induced to mutate in Figures 6A and 6B.

A. Mutation yj41

```
Ce wt          SNERVVLGKGTYGTVYAARDMDTQRQIVVKEIEVKYDEEVQPLMEEISLHSTLCHANIVQ
Ce yj41        SNERVVLEKGTYGTVYAARDMDTQRQIVVKEIEVKYDEEVQPLMEEISLHSTLCHANIVQ

W. BANCROFTI   --DRVVLGRGTYGVVYAGRDLSTQRSIVVKEVEVKNEEEVQPLMEEIQLHSTLSHQNIVQ
L. LOA         SGDRVVLGRGTYGVVYAGRDLSTQRSIVIKEVEVKNEEEVQPLMEEIQLHSTLSHQNIVQ
B. MALAYI      SGDRVVLGRGTYGVVYAGRDLSTQRSIVVKEVEVKNEEEVQPLMEEIQLHSTLSHQNIVQ
A. SUUM        --DRVVLGRGTYGVVYSGRDLTTQRSIVVKEVEVKNEEEVQPLMEEIQLHSTLSHQNIVQ
H. CONTORTUS   --DRVVLGKGTYGVVYSARDVTTQRQIVVKEIEVKYDEEVQPLMEEINLHSTLSHQNIVQ
A. CEYLANICUM SGERAVLGKGTYGVVYSARDVTTQRQIVVKEIEVKYDEEVQPLMEEINLHSTLSHQNIVQ
N. AMERICANUS NGDRVVLGKGTYGVVFSARDVTTQRQIVVKEIEVKYDEEVQPLMEEINLHSTLSHQNIVQ
C. BRIGGSAE    SNERVVLGKGTYGTVYSARDMDTQRQIVVKEVEVKYDEEVQPLMEEIGLHSTLSHQNIVQ
C. REMANEI     SNERVVLGKGTYGTVYSARDMDTQRQIVVKEIEVKYDEEVQPLMEEISLHSTLSHQNIVQ
C. BRENNERI    NNERLVLGKGTYGTVYSARDMDTQRQIVVKEIEVKYDEEVQPLMEEISLHSTLSHQNIVQ
S. RATTI       NNERILLGKGTYGKVFQARDLRTQCSIAVKEIEVKNDDEVQPLMEEIQLHSTLSHRNIVR
T. SUIS        SGQRVVLGKGTFGAVYAGRDRTSQRTIAIKEVQIKNHEEVQPLMEEIQLHSTLAHTNIVQ
T. TRICHIURA   SGQRVVLGKGTFGAVYAGRDRTSQRTIAIKEVQIKNHEEVQPLMEEIQLHSTLAHTNIVQ
T.SPIRALIS     NDQRVVLGKGTFGTVYAGRDLNSQRTIAIKEVEIKNHEEVQPLMEEIQLHSTLVHPNIVQ
T. CANIS       --DRVVLGRGTYGVVYSGRDLTTQSRFNRLSDHECCHAPITVVVSSKNMFFDDSISHPSQ
H. SAPIENS     --ERVVLGKGTYGIVYAGRDLSNQVRIAIKEIPERDSRYSQPLHEEIALHKYLKHRNIVQ
```

B. Mutation yj15

```
Ce wt           LHELKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFT--------GTLQY
Ce yj15         LHELKIVHRNIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFT--------GTLQY

W. BANCROFTI    LHDQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFA--------GTLQY
L. LOA          LHDQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFA--------GTLQY
B. MALAYI       LHDQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFA--------GTLQY
A. SUUM         LHEQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFT--------GTLQY
H. CONTORTUS    LHEQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFT--------GTLQY
A. CEYLANICUM   LHEQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFT--------GTLQY
N. AMERICANUS   LHEQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFT--------GTLQY
C. BRIGGSAE     LHEQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFT--------GTLQY
C. REMANEI      LHDQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFT--------GTLQY
C. BRENNERI     LHEQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFT--------GTLQY
S. RATTI        LHEQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGINPVTDTFK--------GTSQY
T. SUIS         LHSQKIVHRDIKGDNVLVNTYSGLCKIVDFGTCKRLAGLNPVTDTFT--------GTLQY
T. TRICHIURA    LHSQKIVHRDIKGDNVLVNTYSGLCKIVDFGTCKRLAGLNPVTDTFT--------GTLQY
T.SPIRALIS      LHSQKIVHRDIKGDNVLVNTYSGQCKISDFGTCKRLAGLNPIADTFTGLCSPVVVGTLQY
T. CANIS        LHEQKIVHRDIKGDNVLVNTYSGVCKISDFGTCKRLAGLNPVTETFT--------GTLQY
H. SAPIENS      LHENQIVHRDIKGDNVLVNTYSGVVKISDFGTSKRLAGVNPCTETFT--------GTLQY
```

Figure 7: Location of mutations in the catalytic domain of NSY-1 compared to the sequence of the other nematode species present in alignment as well as *H. sapiens*. The amino acid that was induced to mutate in *nsy-1* is colored red. 7A) Mutation *yj41* which when not mutated shows a uniform identity of glycine among all species included in the catalytic site alignment. In *yj41*, the mutant amino acid is glutamic acid. 7B) Mutation *y15* has the amino acid identity of aspartic acid when not mutated among all species included in the catalytic site alignment. The mutated amino acid in *yj15* is arginine.

Mutations *yj41* and *yj15* as seen in Figures 7A and 7B show the amino acids induced to mutate located in the catalytic domain. Both amino acids are the same for all species being

compared. The amino acid mutated for mutation *yj41* is glycine (G) (Figure 7A), converted to a mutant glutamate. This represents a change from a nonpolar amino acid to a negatively charged amino acid residue. The wild type amino acid for mutation *yj15* is aspartic acid, converted to asparagine in the mutant. This is the same change as for *yj13,* with a negative charge replaced by a positive charge. The surrounding sequence for these locations also shows high level of similarity between the species being compared.

Mutation ky397

```
Ce wt           TISDESSNSSSRFFMLQKDSERRRSLGQFMQDYKDLIIDSWSTLLIKQSD----TELVVT
Ce ky397        TISDESSNSSSRFFML*KDSERRRSLGQFMQDYKDLIIDSWSTLLIKQSD----TELVVT

A. SUUM         TSSDESGMSS-RFFMLKKDSERRNTLARFMLEYKAQIIDNWFEHLAKNQQT---TELLVT
T. CANIS        TSSDESGMSS-RFFMLKKDSERRNTLARFMQEYKAEIIDHWFEHLAKSQQT---NELLVT
L. LOA          TASDESGMSS-RFFMLKKDSERRNTLARFMLDYKIEIIDNWYEHLTKSQLT---GDLVVT
C. BRIGGSAE     TISDESSNSSSRFFMLQKDSERRRSLGQFMTDYKDLIIDSWSTLLIQQSD----TELVVT
C. BRENNERI     TISDESSNSSNRFFMLQKDSERRRSLGAFMQDYKDLIIDSWTTLLIQQSD----TELNVT
C. REMANEI      TISDESSNSSNRFFMLQKDSERRRSLGQFMQDYKDLIIDSWSTLLIKQSD----TELVVT
H. CONTORTUS    TISDESANSSNRFFMLKKDSERRHTLAQFMSDYSEQIIDTWMRGIIASMDS---AEVVVT
S. RATTI        LSYSDEHNSTNRFFLLQKDSERRVYLSDFMEKHDTEIIKKWRANLNVLSEGDTSPTVDEN
H. SAPIENS      GLASSPEDRDQGLFLLRKDSERRAILYKILWEEQNQVASNLQECVAQSSE-----ELHLS
```

Figure 8: Location of the mutation *ky397* in the C terminal of *C. elegans nsy-1* protein sequence which when not mutated has the amino acid identity of glutamine for species part of the Caenorhabditis genus. The amino acid that was induced to mutate is colored red.

In *nsy-1*'s C terminal domain there is only one mutation, *ky397.* For *C. elegans* the affected amino acid is glutamine, which is converted to a stop codon in *ky397.* This is consistent with the other *Caenorhabditis* species, as well as *S. ratti*. The other amino acid represented at this position is lysine, and it is seen in the protein sequences for *A. suum, T. canis, L. loa,* and *H. contortus.* The amino acid at this position for *H. sapiens* is arginine (R) (Figure 8). The identity of the amino acids at this mutation spot are not uniform, but at the same time there are no gaps as seen in Figure 6B and 6C.

Coiled-Coil Domains

ASK-1 has coiled-coil domains residing in both the N terminal domain (amino acids 290-317) and the C terminal domain amino acids 1236-1293 (reviewed in Hayakawa et al 2006). By searching for coiled-coil domains in NSY-1, I found a coiled-coil domain in the C terminal domain. According to a coiled-coil region predictor (Lupas et al et al, 1991) the region where NSY-1's coiled-coil domain resides in the *C. elegans* sequence are not in regions of high similarity to any of the other species. However, upon the majority of the species in our set contain a coiled-coil domain in their C terminal domain, similar to *C. elegans*. Of the species that are compared in the phylogenetic trees, including *H. sapiens*, the species that either have no coiled-coil domain or have such a motif in a different part of the protein are *A. ceylanicum*, *W. bancrofti*, *T. spiralis,* and *N. americanus* (Supp Material 2). *A. ceylanicum* is predicted to

have a coiled-coil domain in the N terminal domain (Supp Material 2J) while *W. bancrofti* (Supp Material 2M), *T. spiralis* (Supp Material 2N), and *N. americanus* (Supp Material, 2O) show no probability that a coiled-coil domain exists in their sequences. Also *C. brenneri* appears to have a high likelihood of two coil-coiled domains in its sequence (Supp Material, 2B) (Lupas et al, 1991). This information should be considered when trying to find an inhibitor to the parasitic species equivalent to *nsy-1.*

Discussion

Concerning the future in this line of research, there are two ways of looking at the results, with the first point being the NSY-1-like sequences that are most closely related to *C. elegans* NSY-1. From Figure 4A-4C we see that the sequences from nematode order Spirurida are the most closely related to those of the genus *Caenorhabditis*. This includes the species *A. ceylanicum*, *H. contortus*, and *N. americanus*, if we are focusing on the catalytic site (Figure 4B). Of these three species, if the N terminal is to be the area where the NSY-1 is targeted for inhibition, then *A. ceylanicum* would be the best nematode species to begin attempting to inhibit NSY-1. At the same time, if the C terminal becomes the sequence section of focus then *H. contortus* would be the species that is the closest related to the genus Caenorhabditis and probably the best chance at beginning inhibition of NSY-1 in a parasitic nematode species. At the same time, in the phylogenetic tree that contains all three species of order Spirurida (the catalytic site), *N. americanus* is the nematode species most closely related to the Caenorhabditis *nsy-1* gene.

Along with placement on the phylogenetic tree in relation to *C. elegans* and common motifs another way to determine the species that will be likely to respond best to therapeutic treatment through NSY-1 inhibition is to look at the induced mutations. By comparing each species sequence at the location of these mutations there are two questions to ask: 1. How consistent is the similarity between parasitic species? ; and 2. Would possible therapeutic treatment also affect ASK-1 protein in *H. sapiens*?. If we were to create an inhibitor of the kinase domain it would prove to also inhibit the actions of ASK-1. We can see this from the mutations in the Catalytic domain (Figure 7, mutations *yj41* and *yj15*, since identity of the amino acid that is induced is the same for all of the species involved, including *H. sapiens*. For *yj41* that amino acid is glycine and for *yj15* the amino acid is aspartic acid.

A way to find compounds that inhibit parasitic NSY-1 homologs is by high-throughput screening of a library of chemical compounds. High-throughput screens could be devised using the Nsy-1 or Srf-6 mutant phenotype. If exposure to an inhibitor phenocopies the Nsy-1 or Srf-6 phenotype, it could be because the inhibitor interacts with NSY-1/SRF-6. In that case, the detailed information provided by the present project would be useful in understanding how the inhibitor works, and whether it would be applicable to parasitic nematodes. For example, the similarity of parasitic *nsy-1/srf-6* –like proteins in parasitic nematodes, analyzed here, might point to which species would be most susceptible to inhibition.

Acknowledgements

I would like to thank Professor Samuel Politz for his council in this project as well as Professor Luis Vidali and his lab for allowing me to use their copy of the Geneious program.

References

Bethony J, Brooker S, Albonico M, Geiger S, Loukas A, Diemert D, Hotez P. (2006) Soil-transmitted helminth infections: ascariasis, trichuiasis, and hookworm. *The Lancet* 367: 1521-1532.

Blaxter M. (2011). Nematodes: The Worm and Its Relatives. *PLos Biol*. 9. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001050

Blaxter M, De Ley P, Liu L, Scheldeman P, Vierstraete A, Vanfleteren J, Mackey L, Dorris M, Frisse L, Vida J, Thomas W. (1998). A molecular Evolutionary Framework for the Phylum Nematoda. *Nature* 392: 71-75.

CDC: Centers for Disease Control and Prevention. (2012). Parasites-Trichinellosis. http://www.cdc.gov/parasites/trichinellosis/.

CDC: Centers for Disease Control and Prevention. (2013) Parasites-Lymphatic Filarsis. http://www.cdc.gov/parasites/lymphaticfilariasis/gen_info/faqs.html.

CDC: Centers for Disease Control and Prevention. (2015). Loiasis. http://www.cdc.gov/parasites/loiasis/epi.html.

Cox, G. N., Staprans, S., & Edgar, R. S. (1981). The cuticle of Caenorhabditis elegans: II. Stage-specific changes in ultrastructure and protein composition during postembryonic development. *Developmental biology, 86*: 456-470.

Edgar R, Cox G, Kusch M, Politz J. (1982). The Cuticle of Caenorhbaditis elegans. *Journal of Nematology* 14: 248-258

De Ley P. (2006). A Quick Tour of Nematode Diversity and the Backbone of Nematode Phylogeny. *Wormbook*: 1-8. http://www.wormbook.org/chapters/www_quicktourdiversity/quicktourdiversity.html

Dembo, A., Karlin, S. & Zeitouni, O. (1994) Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob*. 22: 2022 Gumbel, E. (1958). Statistics of extremes. Columbia University Press, New York, NY.

Grenache D, Caldicott I, Albert P, Riddle D, Politz S. (1996) environmental induction and genetic control of surface antigen switching in the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci* 93: 12388-12393

Hayakawa Y, Hirata Y, Nakagawa H, Sakamoto K, Hibiba Y, Kinoshita H, Nakata W, Takahashi R, Tateishi K, Tada M, Akanuma M, Yoshida H, Takeda K, Ichijo H, Omata M, Maeda S, Koike K. (2006). Apoptosis signal-regulating kinase 1 and cyclin D1 compose a positive feedback loop contributing to tumor growth in gastric cancer. *PNAS*. 108: 780-785.

Hemmer R, Donkin S, chin K, Grenache D, Bhatt H & M, Politz S. (1991) Altered expression of an L-1-specific, O-linked cuticle surface glycoprotein in mutants of the nematode *Caenorhabditis elegans. J. Cell Biol* 115: 1237-1247

Henikoff S and Henikoff J. (1992). Amino acid substitution matrices from protein blocks. *PNAS*. 89: 10915-10919

Karlin, S. & Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci* 87:2264-2268

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*: 1647-1649.

Lupas et al A, Van Dyke M, Stock J. (1991). Predicting Coiled Coils from Protein Sequences. Science. 252: 116-1164.

Olsen D, Phu D, Libby L, Cormier J, Montez K, Ryder E, Politz S. (2006) Chemosensory control of surface antigen switching in the nematode *Caenorhabditis elegans*. *Genes, Brain and Behavior* 6: 240-252.

Riddle D, Blumenthal T, Meyer B, Priess J. (1997). C. elegans II. 2nd Edition. Cold Spring Harbor (NY) Cold Spring Harbor Laboratory Press. Available from: http://www.ncbi.nlm.nih.gov/books/NBK19997/

Sagasti A, Hisamoto N, Hyodo J, Tanaka-Hino M, Matsumoto K, C Bargmann. (2001). The CaMKII UNC-43 Activates the MAPKKK NSY-1 to Execute a Lateral Signaling Decision Required for Asymmetric Olfactory Neuron Fates. *Cell* 105: 221-232.

Sellers P. (1984). Pattern recognition in genetic sequences by mismatch density. *Bull. Math. Biol* 46: 501-514.

Tanaka-Hino M, Sagasti A, Hisamoto N, Kawasaki M, Nakano S, Ninomiya-Tsuji J, Bargmann C, Matsumoto K. (2002). SEK-1 MAPKK mediates Ca^{2+} signaling to determine neuronal asymmetric development in *Caenorhabditis elegans*. *EMBO reports* 3: 56-62.

Trichinella.org. (2004). Life Cycle of *Trichinella spiralis*. http://www.trichinella.org/bio_lifecycle.htm

Troemel, E. (1999). Chemosensory signaling in *C. elegans*. *BioEssays* 21: 1011-1020.

Uniprot.(2015). Caenorhabditis. Taxonomy. http://www.uniprot.org/taxonomy/6237

Uniprot. (2015). Rhabditida. Taxonomy. http://www.uniprot.org/taxonomy/6236

Uniprot. (2015). Strongylida. Taxonomy. http://www.uniprot.org/taxonomy/6308

Uniprot. (2015). Spirurida. Taxonomy. http://www.uniprot.org/taxonomy/6274

Uniprot. (2015). Trichocephalida. Taxonomy. http://www.uniprot.org/taxonomy/6329

Van de Peer Y, Maere S, Meyer A. (2009). The Evolutionary Significance of Ancient Genome Duplications. *Nature Reviews Genetics* 10: 725-732

Wes P and Bargmann C. (2001). *C. elegans* ordour discrimination requires asymmetric diversity in olfactory neurons. *Nature* 410: 698-701

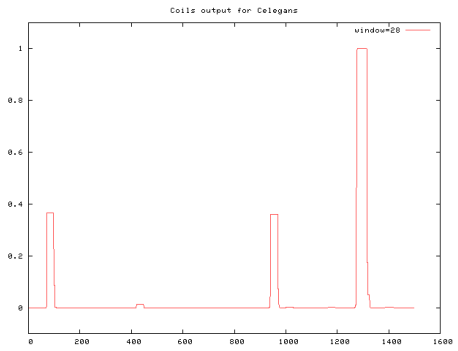
Supplementary Material

1. Step-by-Step method to the final Phylogenetic Tree product including the algorithms being used for each step.

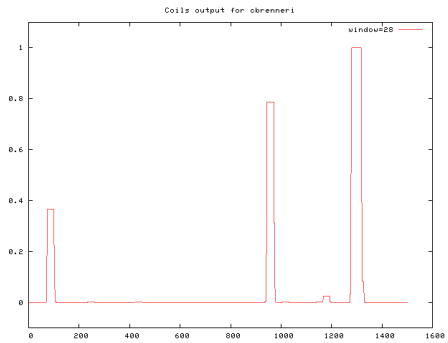
    A. BLAST done on each domain of the *nsy-1* query. Each domain being the N terminal, Catalytic domain, and the C terminal
        a. Species selected based on:
            i. E-value (close to zero as possible)
            ii. Percentage of similarity (as close to 100% as possible)
            iii. Whether the species was parasitic or not.
                1. The only species that were not parasitic were those that were part of the Caenorhabditis genus.
    B. Species chosen were aligned with the *nsy-1* query
        a. The alignments were made using the CLUSTALW algorithm
    C. Phylogenetic trees were made from these alignments using the two methods:
        a. Neighbor-joining method
        b. Jukes-Cantor alignment

2. Coil-coiled domain predictions for the sequences from each species being compared to our query. The probability of a coil-coiled domain existing in a section of sequence is on the y-axis, and the position within the sequence is shown by the x-axis. Any probability of sixty or higher means it is safe to say there is a coil-coiled domain present at that given location.

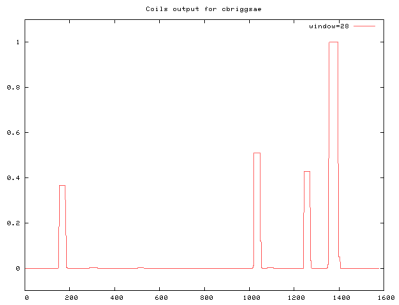
A. *C. elegans*



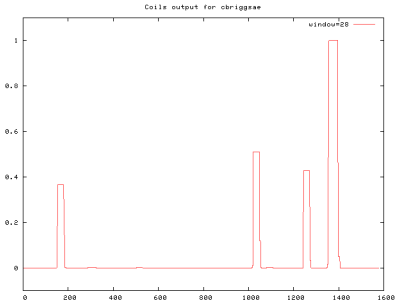
B. *C. brenneri*



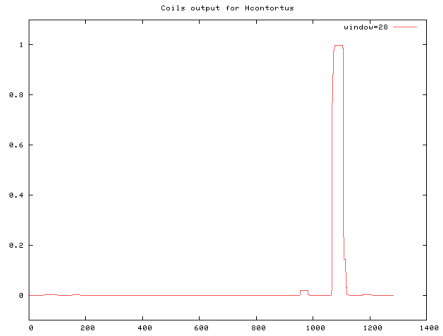
A. *C. briggsae*



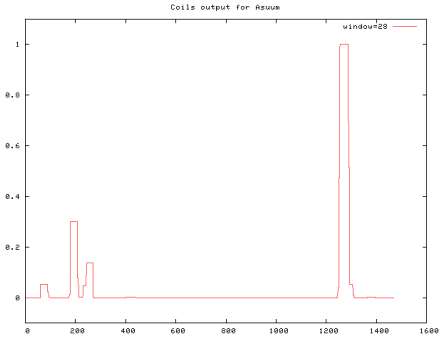
D. *C. remanei*



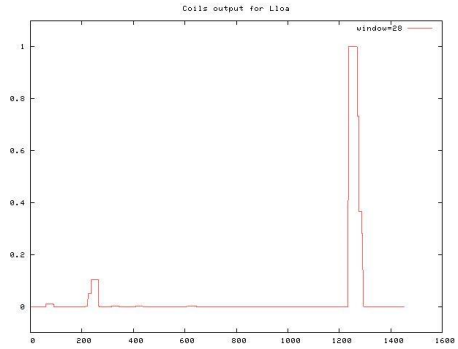
E. *H. contortus*



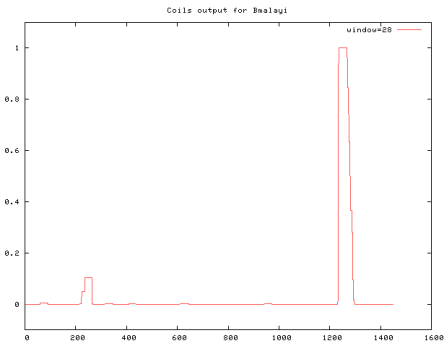
F. *A. suum*



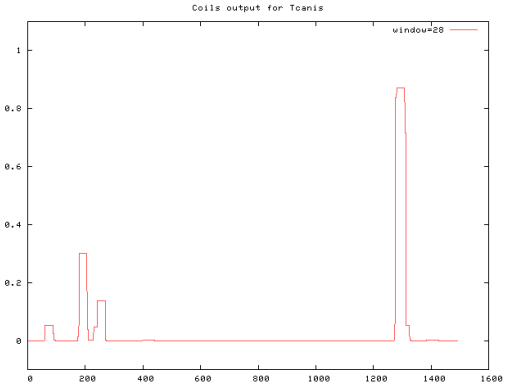
G. *L. loa*



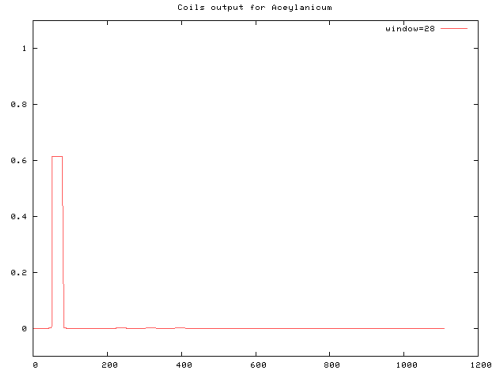
H. *B. malayi*



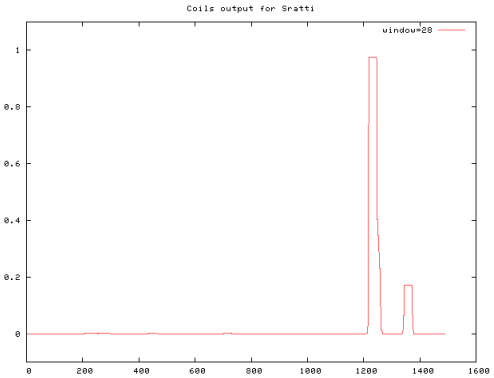
I. *T. canis*



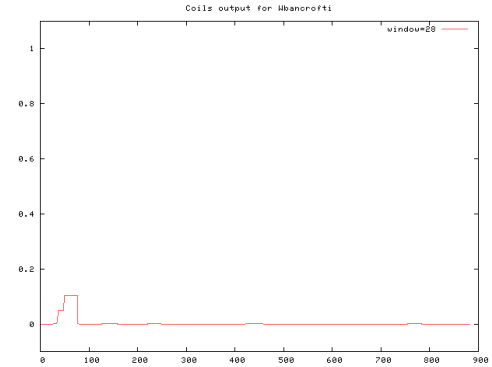
J. *A. ceylanicum*



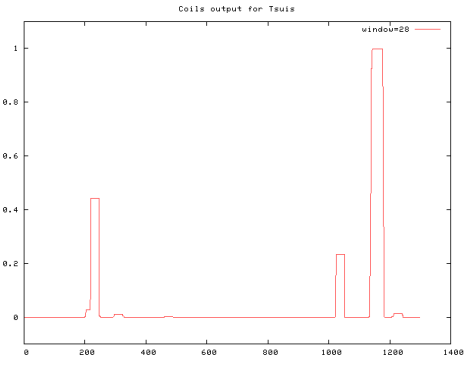
J. *S. ratti*



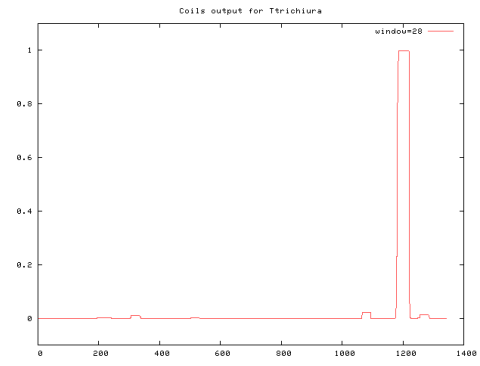
K. *W. bancrofti*



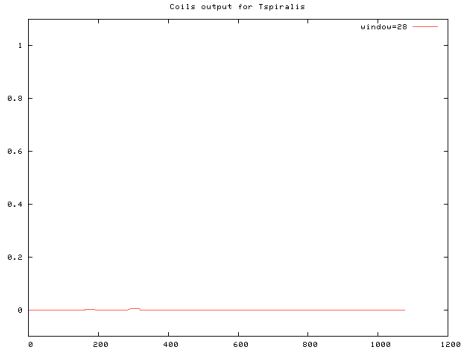
L. *T. suis*



M. *T. trichiura*



N. *T. spiralis*



O. *N. americanus*